



# 数据挖掘技术概述及前景展望

**Data Mining and Prospect**

商业智能研讨沙龙—上海站  
ITPUB ChinaUnix IXPUB 主办

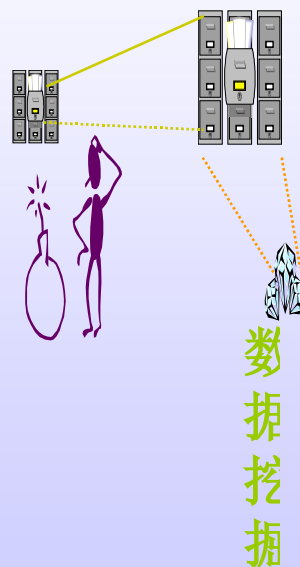
# 数据挖掘由来

- 产生背景

- ➔ 随着数据库技术的飞速发展，快速增长的海量数据收集、存放在大量数据储存库中
- ➔ 理解他们已经远远超出人的能力
- ➔ 数据坟墓——难得再访问的数据档案
- ➔ 数据爆炸，但知识缺乏
- ➔ 人们被数据淹没，却饥饿于知识

# 数据挖掘的原由

数据存储成本越来越低，数据库越来越大……



可怕的数据

有价值的知

识

商业智能研讨沙龙 - 上海站  
ITPUB ChinaUnix IXPUB 主办

## 网络之后的下一个技术热点

“要学会抛弃信息”

“如何才能不被信息淹没，而是从中及时发现有用的知识、提高信息利用率？”

“需要是发明之母”——数据挖掘：海量数据的自动分析技术

数据开采和知识发现（DMKD）技术应运而生

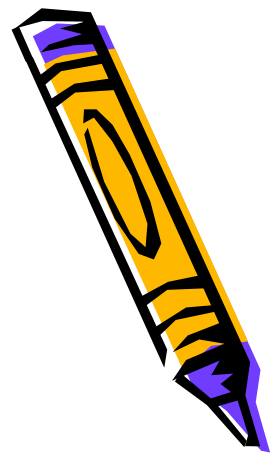
Gartner Group 的一次高级技术调查将数据挖掘和人工智能列为“未来三到五年



# 从商业数据到商业信息的进化

进化阶段	商业问题	支持技术	产品厂家	产品特点
数据搜集 (60年代)	“过去五年中我的总收入是多少？”	计算机, 磁带和磁盘	IBM, CDC	提供历史性的、静态的数据信息
数据访问 (80年代)	“在新英格兰的分部去年三月的销售额是多少？”	关系数据库 (RDBMS), 结构化查询语言 (SQL), ODBC, Oracle, Sybase, Informix, IBM, Microsoft	Oracle, Sybase, Informix, IBM, Microsoft	在记录级提供历史性的、动态数据信息
数据仓库; 决策支持 (90年代)	“在新英格兰的分部去年三月的销售额是多少? 波士顿据此可得出什么结论?”	联机分析处理 (OLAP), 多维数据库, 数据仓库	Pilot, Comshare, Arbor, Cognos, Microstrategy	在各种层次上提供回溯的、动态的数据信息
数据挖掘 (正在流行)	“下个个月波士顿的销售会怎么样? 为什么?”	高级算法, 多处理器计算机, 海量数据库	Pilot, Lockheed, IBM, SGI, 其他初创公司	提供预测性的信息

# 数据挖掘概念的提出



## ■ 现在数据挖掘概念的首次国际学术会议

1989 年 8 月在美国底特律召开的第 11 届国际联合人工智能学术会议 (IJCAI - 89) 上, Gregory Piatetsky-Shapiro 组织了 “数据库中的知识发现” (KDD : Knowledge Discovery in Database) 专题讨论会, 该讨论会的重点是强调发现 (Discovery) 的方法以及发现的是知识 (Knowledge) 两个方面。

## • 相继开展的专题讨论会

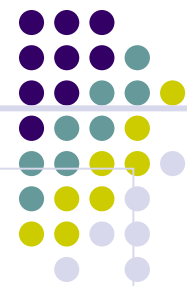
随后在 1991、1993 和 1994 年都举行了 KDD 专题讨论会, 来自各个领域的研究人员和应用开发者集中讨论了数据统计、海量数据分析算法、知识表示和知识运用等问题。



# 数据挖掘概念的提出



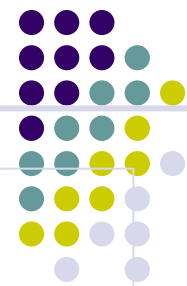
- 第一届 **KDD** 国际学术会议
- 随着参与科研和开发人员的不断增加，国际 **KDD** 组委会于 1995 年把专题讨论会发展成为国际年会。在加拿大的蒙特利尔市召开了第一届 **KDD** 国际学术会。其会议名称全称为 “**ACM SIGKDD** ( **Special Interested Group on Knowledge Discovery in Databases** ) **International Conference on Knowledge Discovery and Data Mining**” 在这次会议上 “数据挖掘” (**Data Mining**) 概念第一次由 **Usama Fayyad** 提出。
- **Usama Fayyad** 对数据挖掘概念的界定  
数据挖掘指的是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、有效的、新颖的、潜在有用的、并且最终可理解的模式的非平凡过程。
- **SAS** 软件研究所对数据挖掘所下的定义是：  
数据挖掘是按照既定的业务目标， 对大量的企业数据进行探索、揭示隐藏其中的规律性并进一步将之模型化的先进、有效的方法。



## 技术上的定义及含义

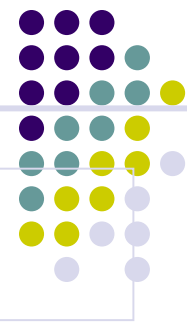
**数据挖掘（Data Mining）**  
就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。





## 技术上的定义及含义

- **数据源必须是真实的、大量的、含噪声的；**
- **发现的是用户感兴趣的知识；**
- **发现的知识要可接受、可理解、可运用；**
- **并不要求发现放之四海皆准的知识，仅支持特定的发现问题**



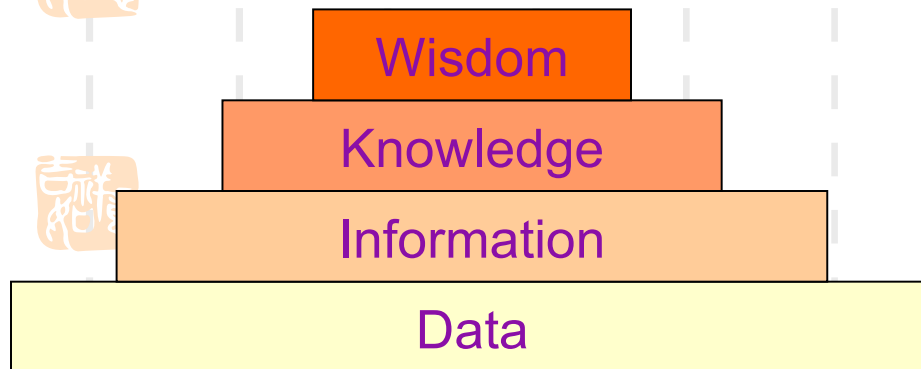
## 商业角度的定义

**数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。**

**按企业既定业务目标，对大量的企业数据进行探索和分析，揭示隐藏的、未知的或验证已知的规律性，并进一步将其模型化的先进有效的方法。**

# 知识是什么... ..

- 知识是对信息进行智能性加工所形成的对客观世界规律性的认识



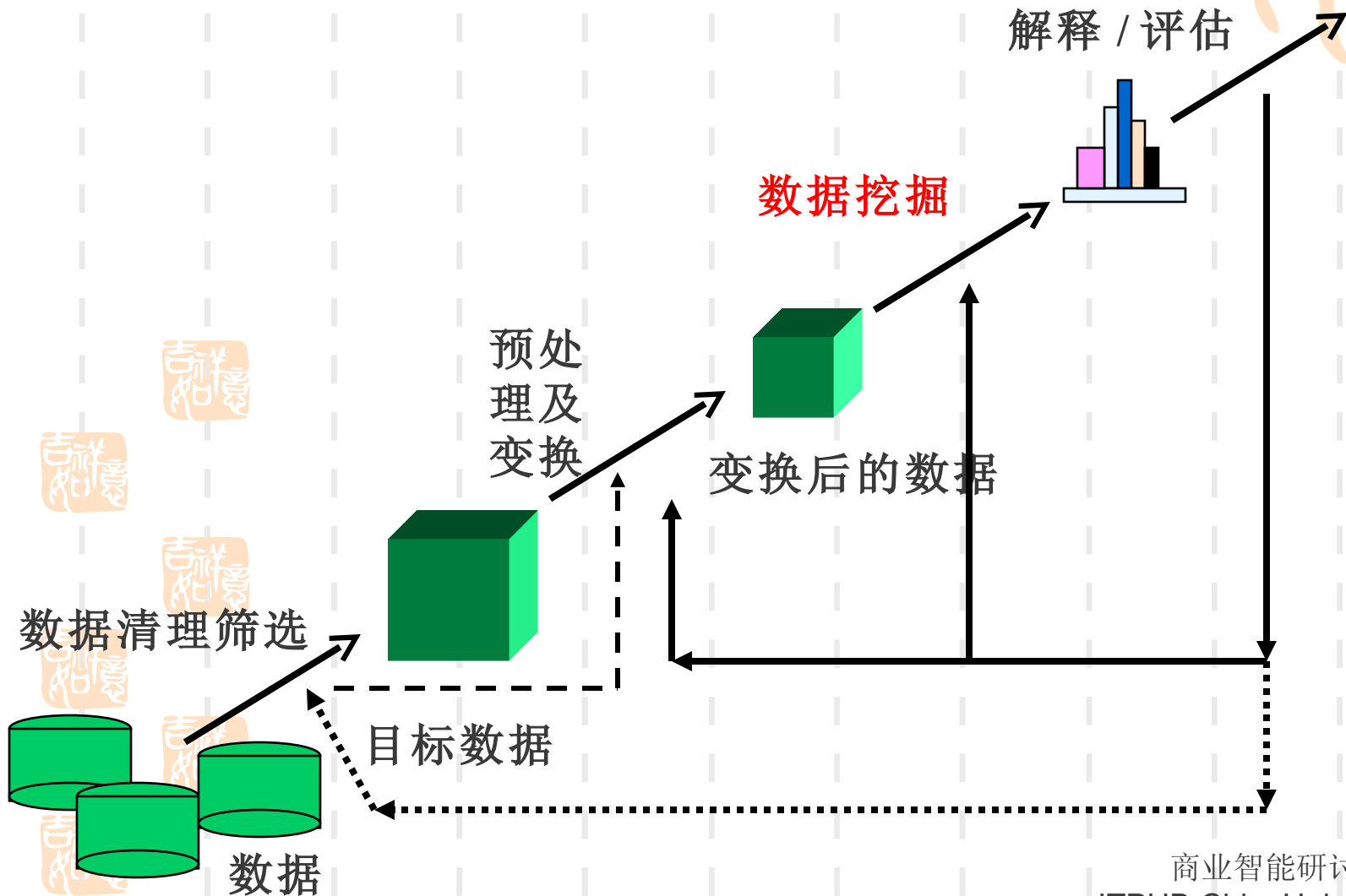
*Knowledge + experience*

*Information + rules*

*Data + context*

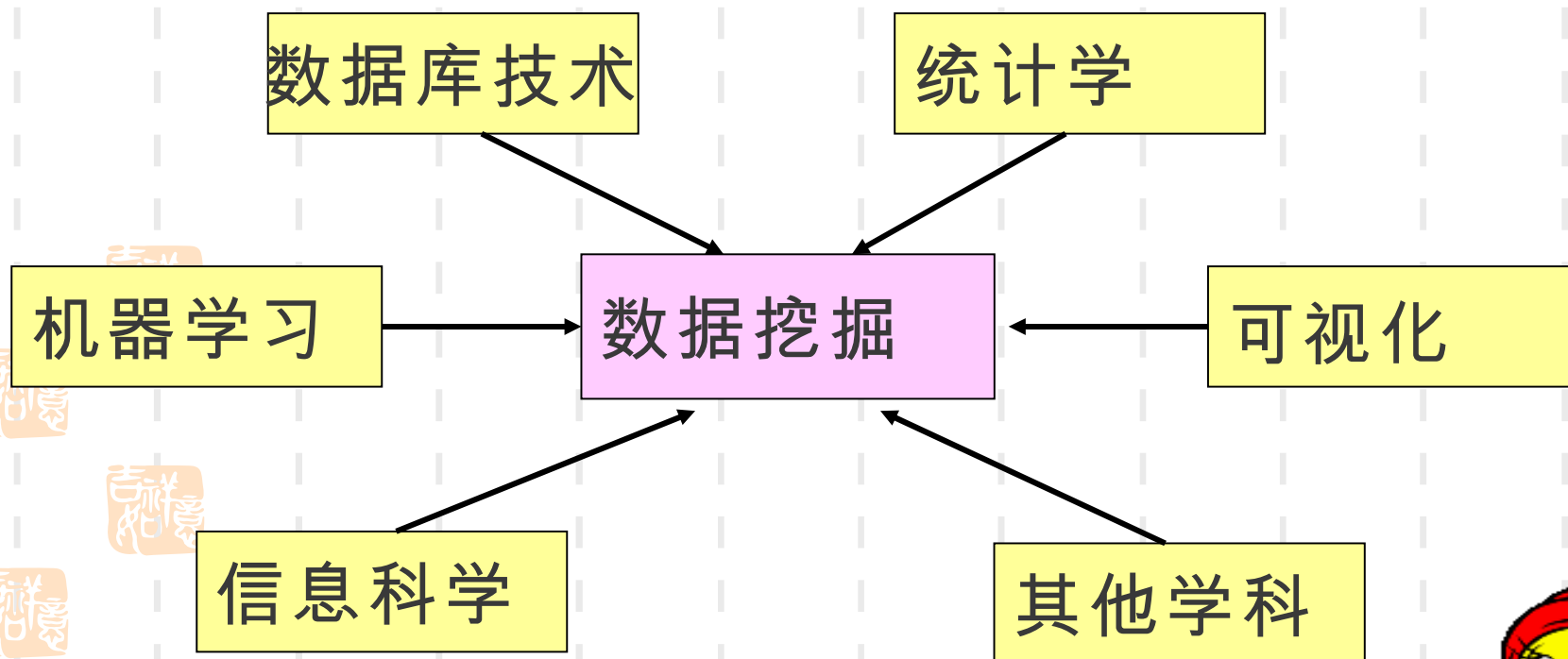
# 知识发现（KDD）的过程

Knowledge

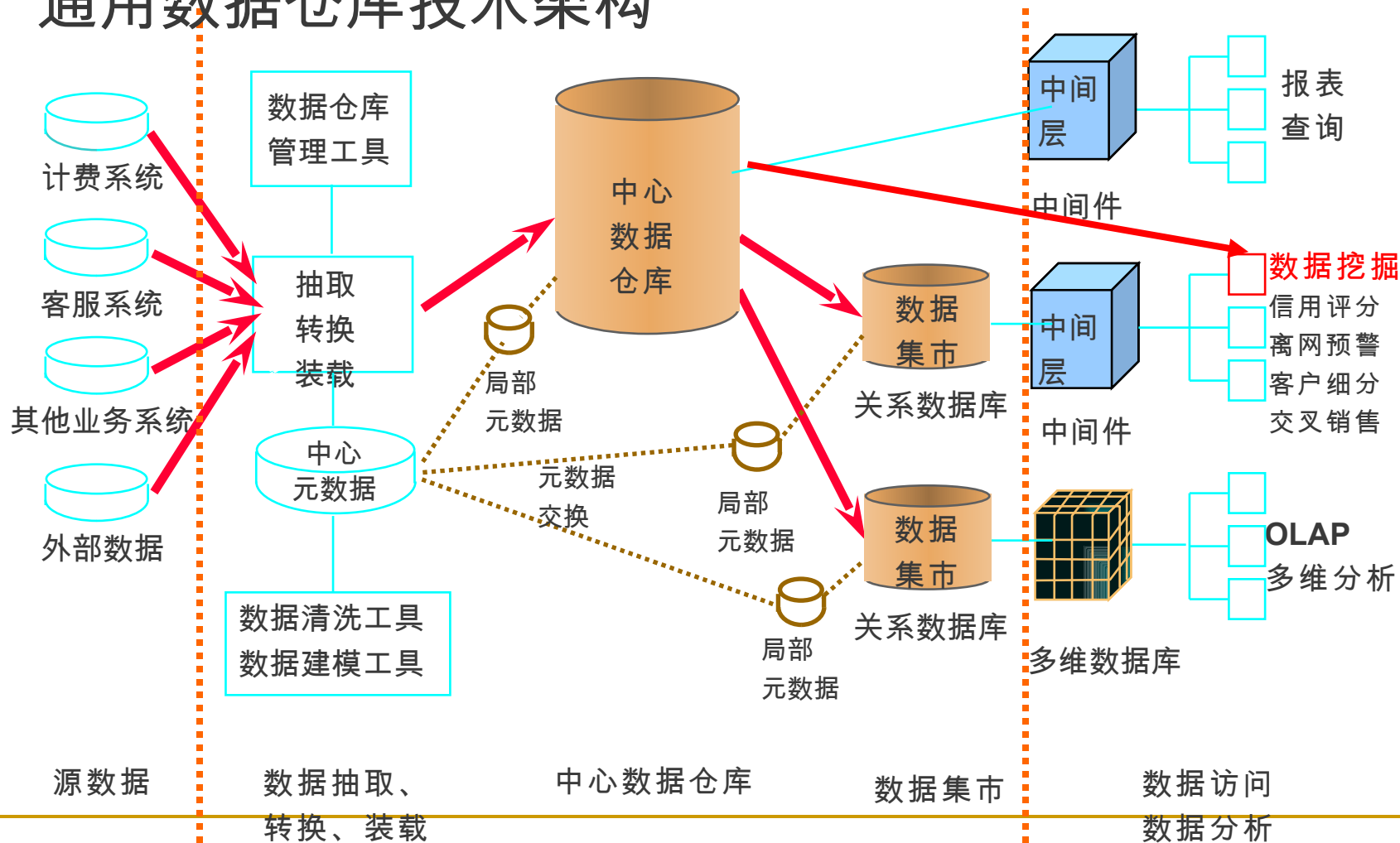




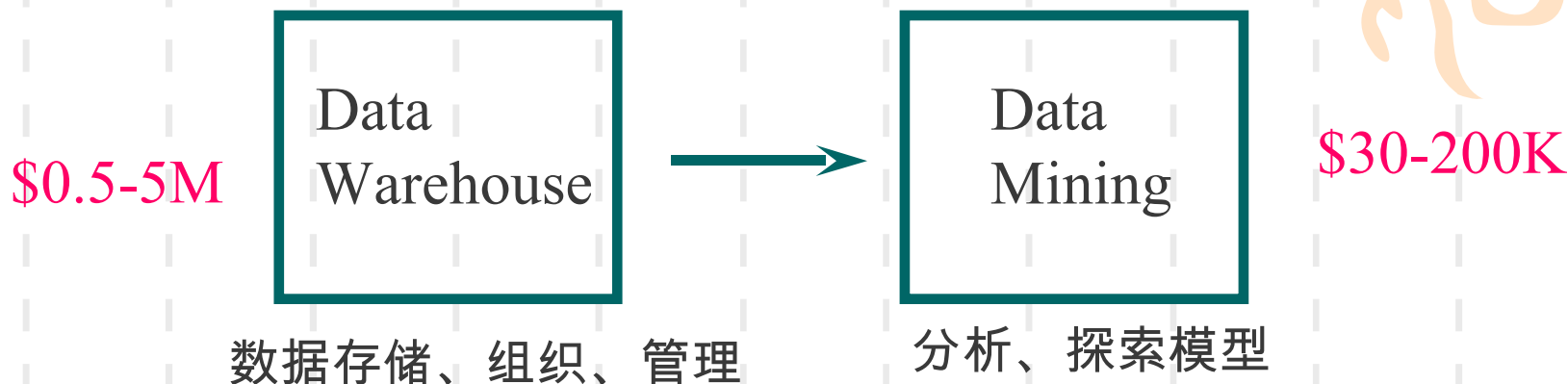
# 数据挖掘是一个交叉学科领域



## 通用数据仓库技术架构



# 数据仓库与数据挖掘的关系



- 数据仓库并不是数据挖掘必需的
- 数据仓库汇总并清理数据，可以作为数据挖掘的基础
  - 数据仓库与数据挖掘都是决策支持新技术。但它们有着完全不同的辅助决策方式。
  - 数据仓库和数据挖掘的结合对支持决策会起更大的作用。

# 数据挖掘与 OLAP



## ■ 数据挖掘与 OLAP 的区别与联系

OLAP 是先建立一系列的假设，然后通过分析来证实或推理这些假设来最终得到自己的结论，本质上是一个演绎推理过程。

数据挖掘是在数据库中自己寻找模型，本质上是一个归纳过程。

两个相辅相成，可以利用 OLAP 验证 DM 的结果。

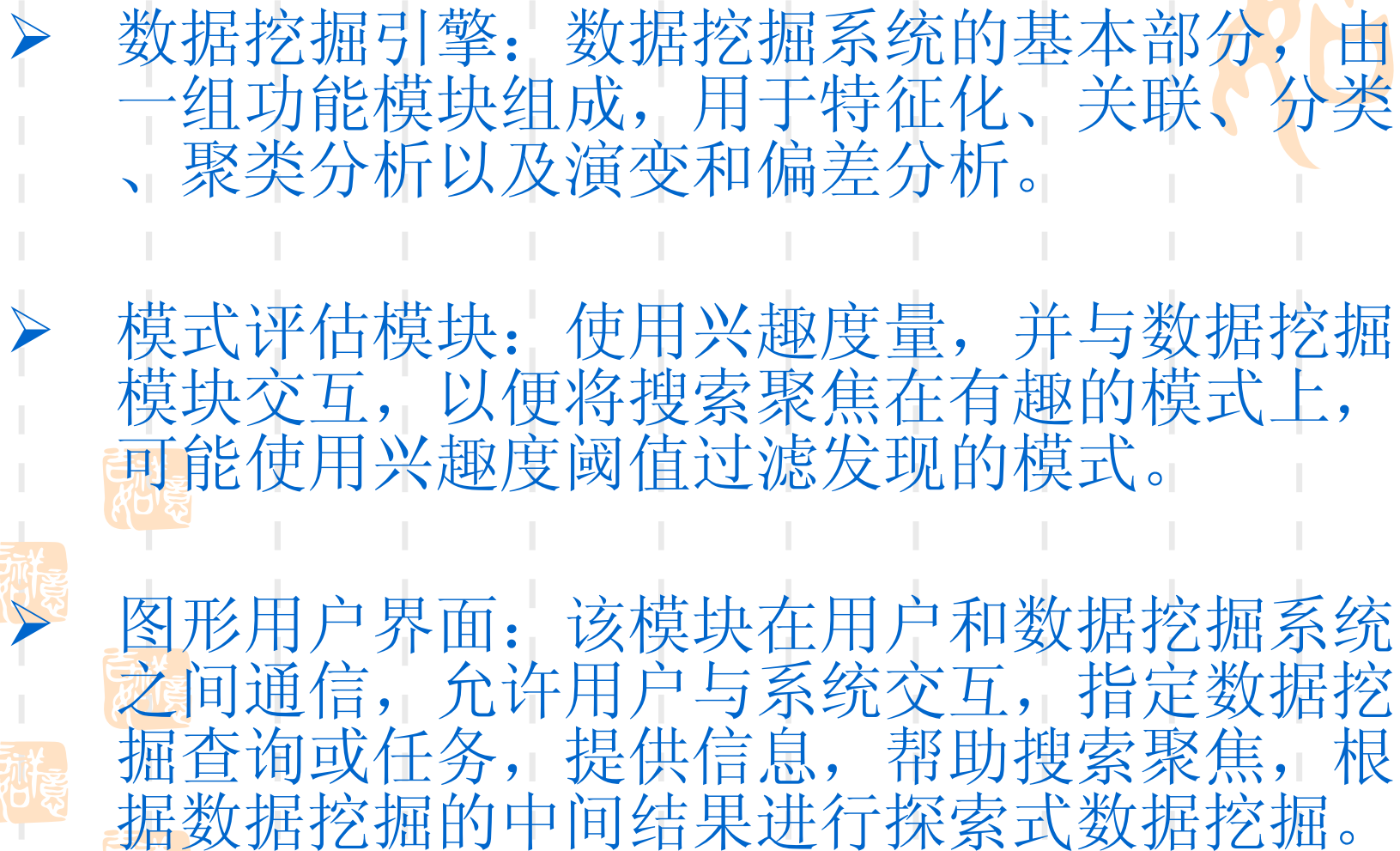
**例：**用数据挖掘工具的分析员想找到引起贷款拖欠的风险因素。然后利用 OLAP 加以验证结论的可靠性。





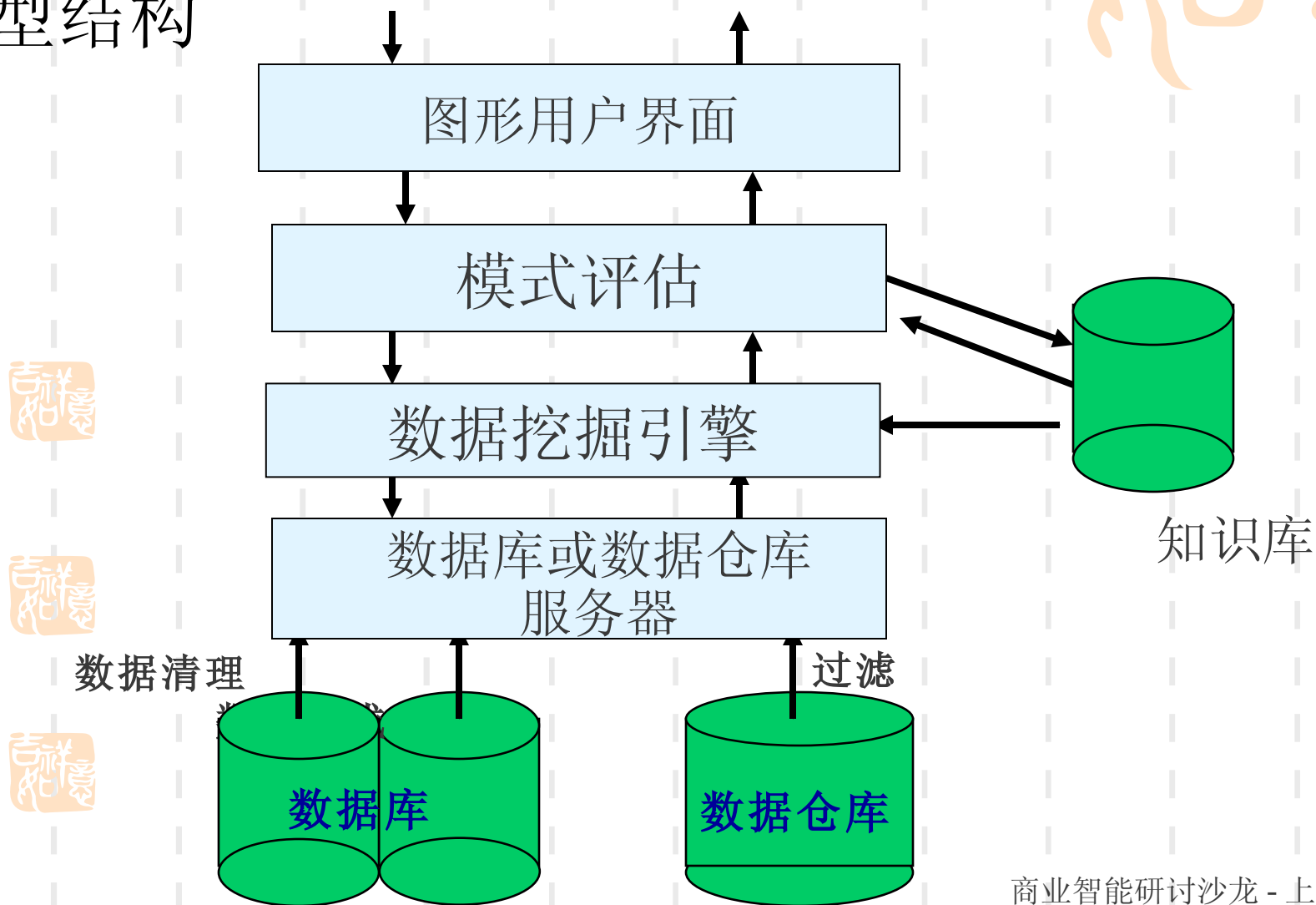
# 数据挖掘系统的组成

- 数据库、数据仓库或其他信息库：是一个或一组数据库、数据仓库、电子表格或其他类型的信息库。可以在数据上进行数据清理和集成。
- 数据库或数据仓库服务器：根据用户的挖掘请求，数据库或数据仓库服务器负责提取相关数据。
- 知识库：是领域知识，用于指导搜索，或评估结果模式的兴趣度。

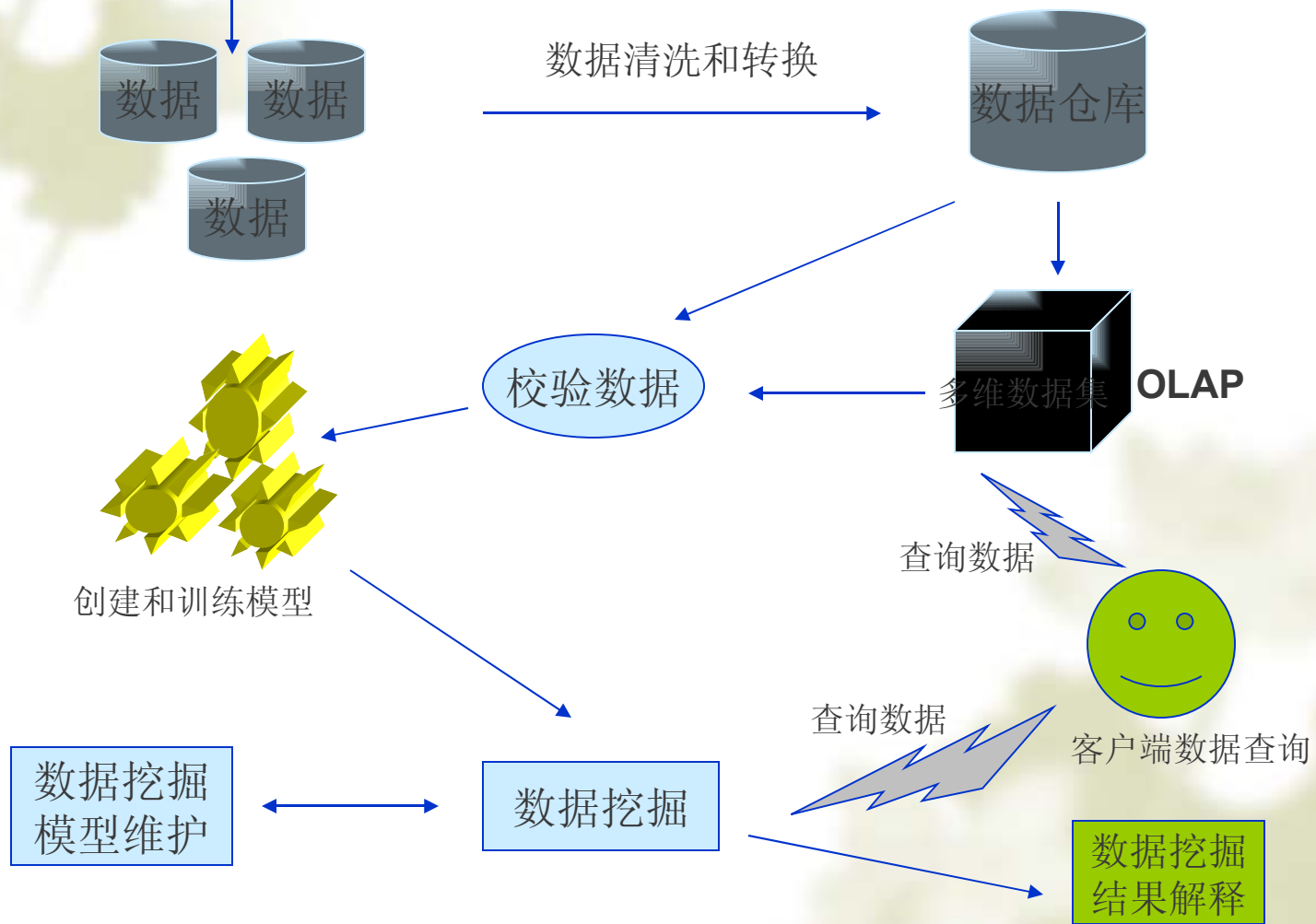
- 
- 数据挖掘引擎：数据挖掘系统的基本部分，由一组功能模块组成，用于特征化、关联、分类、聚类分析以及演变和偏差分析。
  - 模式评估模块：使用兴趣度量，并与数据挖掘模块交互，以便将搜索聚焦在有趣的模式上，可能使用兴趣度阈值过滤发现的模式。
  - 图形用户界面：该模块在用户和数据挖掘系统之间通信，允许用户与系统交互，指定数据挖掘查询或任务，提供信息，帮助搜索聚焦，根据数据挖掘的中间结果进行探索式数据挖掘。

# 数据挖掘系统结构

## 典型结构

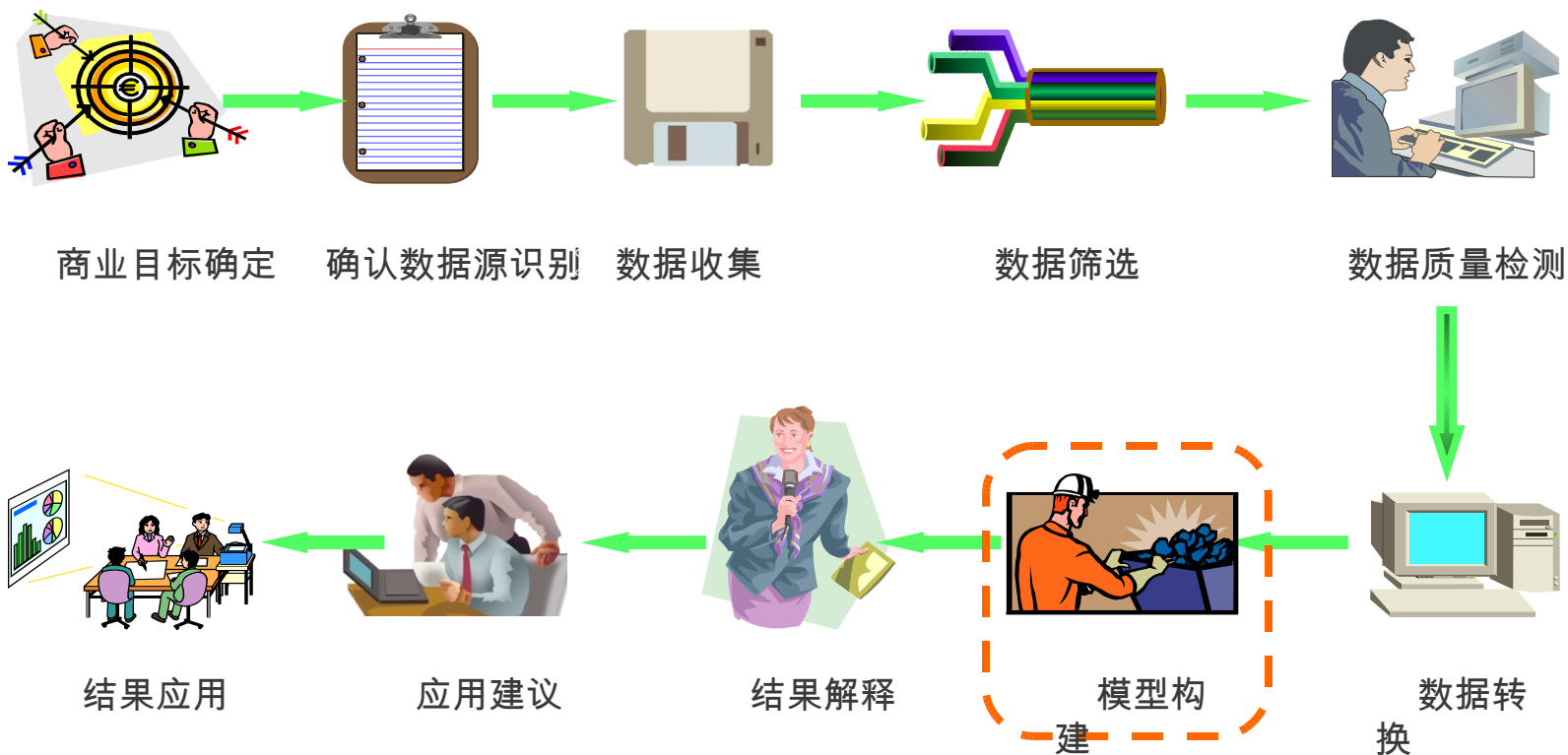


# 数据挖掘过程





# 挖掘项目工作流程



# 数据挖掘过程

- ◆ 数据清理 ( 消除噪声或不一致数据 )
- ◆ 数据集成 ( 多种数据源可以组合在一起 )
- ◆ 数据选择 ( 从数据库中检索与分析任务相关的数据 )
- ◆ 数据变换 ( 数据变换或统一成适合挖掘的形式 )
- ◆ 数据挖掘 ( 使用各种方法提取数据模式 )
- ◆ 模式评估 ( 使用某种度量, 识别真正有趣的模式 )
- ◆ 知识表示 ( 使用可视化和知识表示技术, 向用户提供挖掘的知识 )

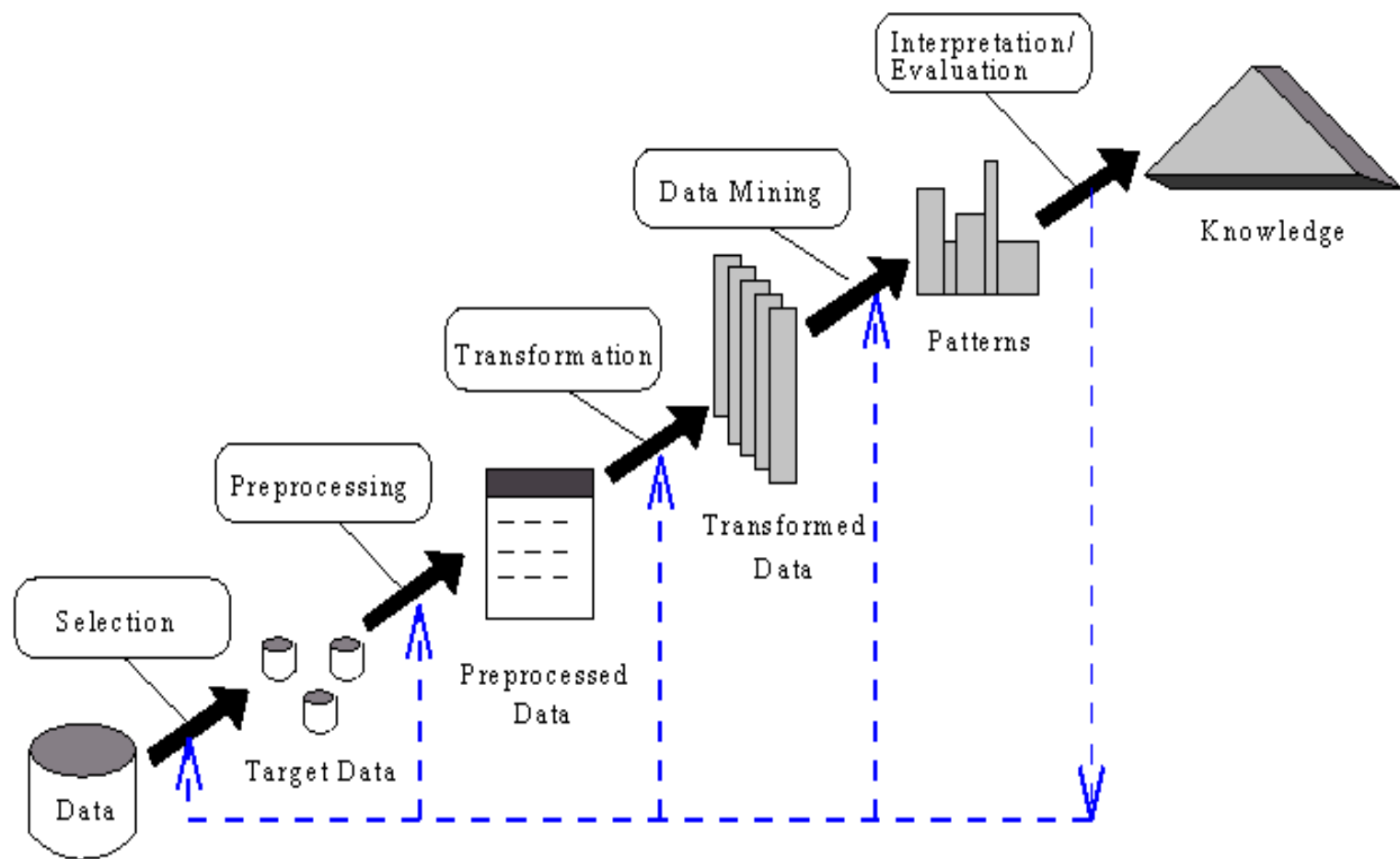
《 data mining concepts and techniques 》



# 从系统设计看数据挖掘过程模型

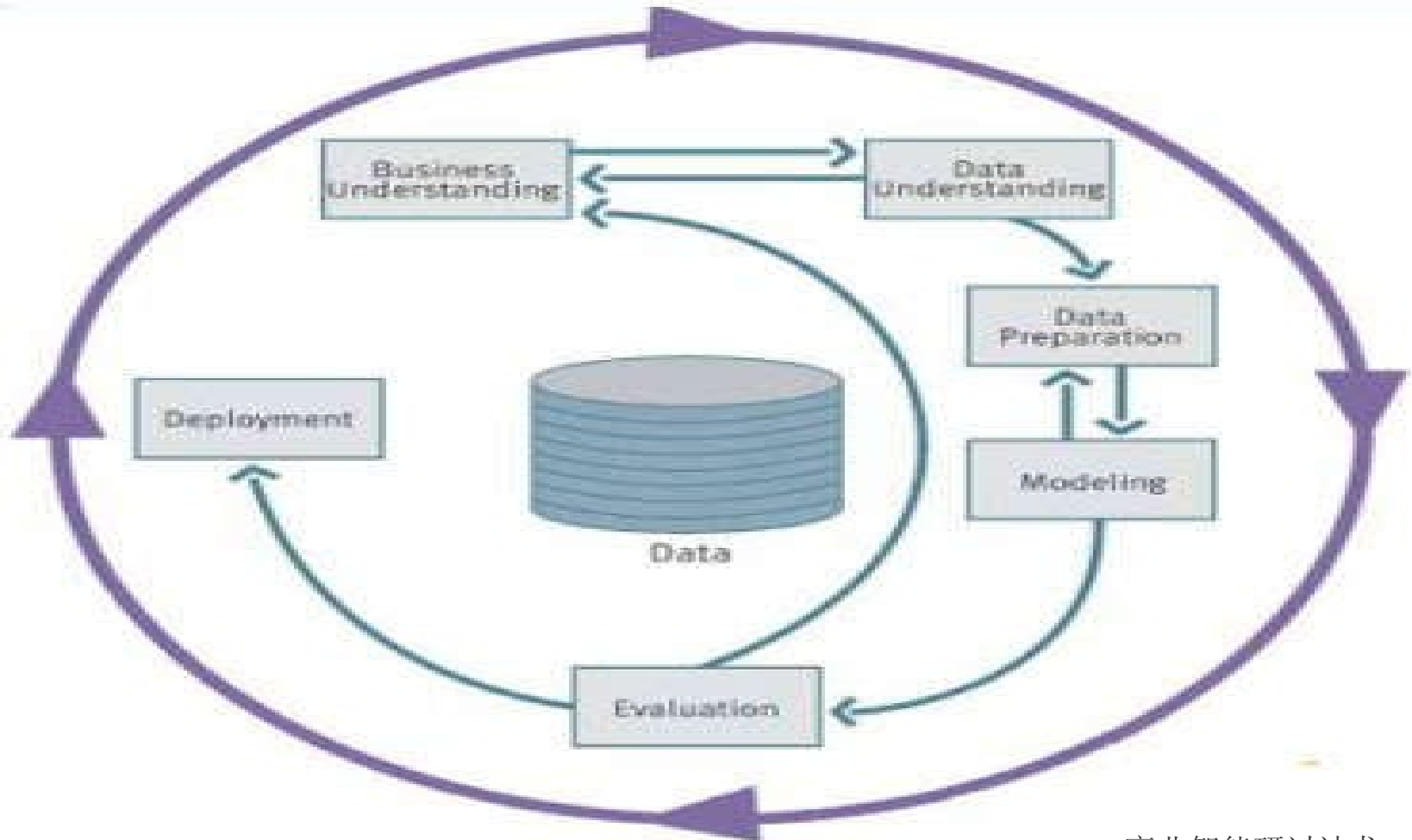
- ◆ 一种是 **Fayyad** 等人总结的过程模型
- ◆ 另一种是遵循 **CRISP-DM** 标准的过程模型

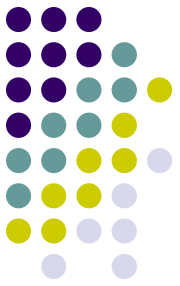
# Fayyad 过程模型





# CRISP -DM ( Cross-Industry Standard Process for Data Mining ) 过程模型

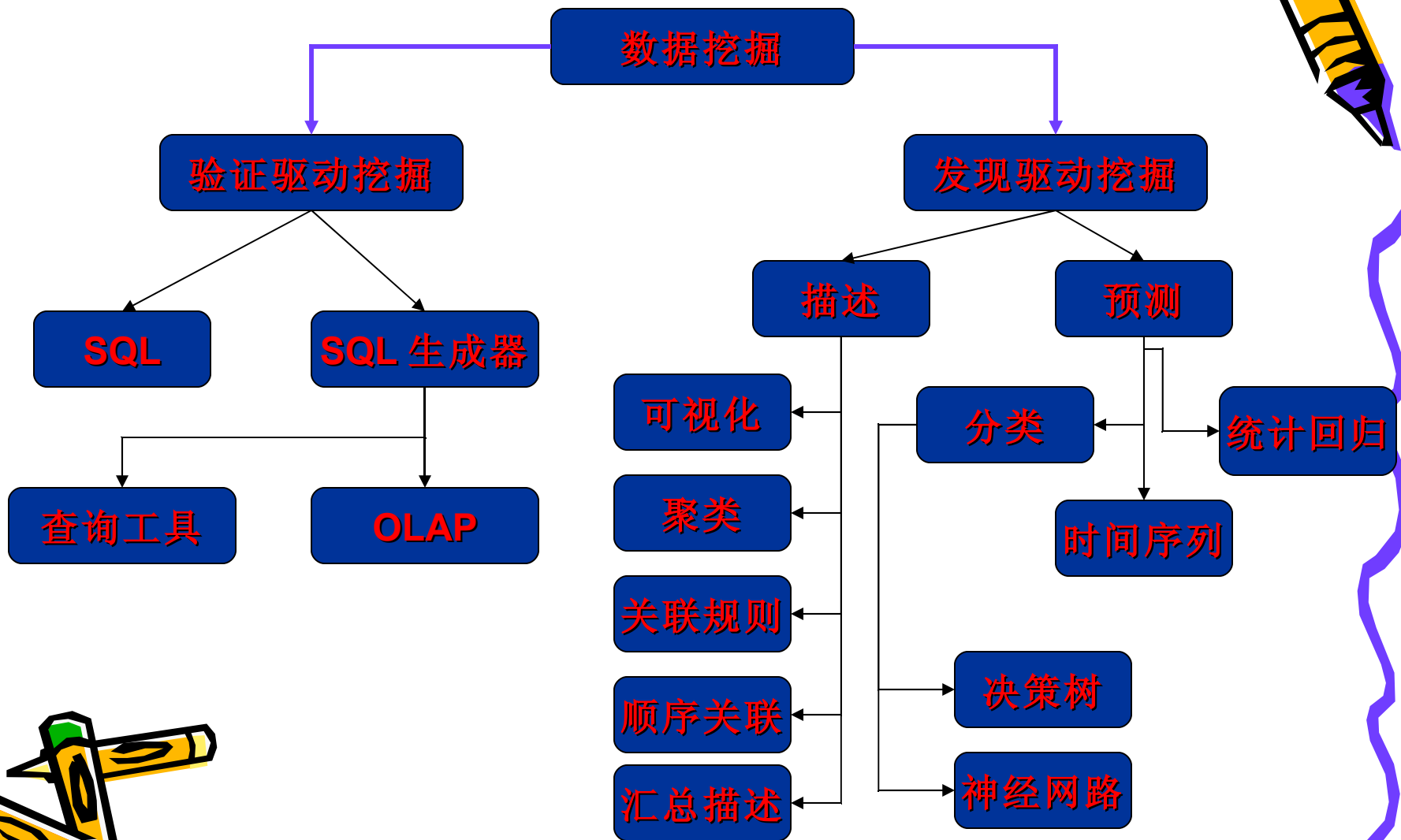




# 数据挖掘方法论

- CRISP\_DM (Cross Industry Standard Process for DM)
  - 1998 年, 由 *NCR*、*Clementine*、*OHRA* 和 *Daimler-Benz* 的联合项目组提出
- SEMMA
  - SAS 公司提出的方法
  - Sample, Explore, Modify, Model, Assess
- 在战略上使用 Crisp\_DM 方法论, 在战术上应用 SEMMA 方法论

# 数据挖掘技术分类



# 数据挖掘的任务和方法

- 数据挖掘的任务是从大量的数据中发现模式。根据数据挖掘的任务可分为多种类型，其中比较典型的有：

- 预测模型
- 关联分析
- 分类分析
- 聚类分析
- 序列分析
- 偏差检测
- 模式相似性挖掘
- **Web 数据挖掘**

# 预测模型



- 预测模型（ **Predictive Modeling** ）：所谓预测即从数据库或数据仓库中已知的数据推测未知的数据或对象集中某些属性的值分布。
- 建立预测模型的常用方法：
  - 回归分析
  - 线性模型
  - 关联规则
  - 决策树预测
  - 遗传算法
  - 神经网络





# 关联分析



- 关联（**Association**）分析：关联规则描述了一组数据项之间的密切度或关系。关联分析用于发现项目集之间的关联。在关联规则挖掘算法中，通常给出了置信度和支持度两个概念，对于置信度和支持度均大于给定阈值的规则称为强规则，而关联分析主要就是对强规则的挖掘。
- 关联分析算法：  
**APRIORI 算法**、**DHP 算法**、**DIC 算法**、**PARTITION 算法**及它们的各种改进算法等。



# 分类分析



- 分类（**Classification**）分析：所谓分类是根据数据的特征为每个类别建立一个模型，根据数据的属性将数据分配到不同的组中。
- 分类分析的常用方法：
  - 粗糙（**Rough**）集
  - 决策树
  - 神经网络
  - 统计分析法



# 聚类分析



- **聚类 (Clustering) 分析：** 所谓聚类是指一组彼此间非常“相似”的数据对象的集合。相似的程度可以通过距离函数来表示，由用户或专家指定。

- **聚类分析的常用方法：**

- 随机搜索聚类法

- 特征聚类

- CF 树 ( 聚类特征数 )



# 序列分析



- 序列（ **Sequence** ）分析：序列分析主要用于分析数据仓库中的某类与时间相关的数据，搜索类似的序列或子序列，并挖掘时序模式、周期性、趋势和偏离等。
- 序列模式可以看成是一种特定的关联模型，它在关联模型中增加了时间属性。
- 例如：在所有购买了彩色电视机的人中，有 **60%** 的人再购买 **VCD** 产品



# 偏差检测

- 偏差检测（**Deviation Detection**）：用于检测并解释数据分类的偏差，它有助于滤掉知识发现引擎所抽取的无关信息，也可滤掉那些不合适的数据，同时可产生新的关注性事实。
- 偏差包括很多有用的知识，如以下 4 类：
  - 分类中的反常实例；
  - 模式的例外；
  - 观察结果对模型预测的偏差；
  - 量值随时间的变化。

# 模式相似性挖掘

- 模式相似性挖掘：用于在时间数据库或空间数据库中搜索相似模式时，从所有对象中找出用户定义范围内的对象；或找出所有元素对，元素对中两者的距离小于用户定义的距离范围。

模式相似性挖掘的方法有相似度测量法、遗传算法等。

# Web 数据挖掘



- **Web 数据挖掘：**万维网是一个巨大的、分布广泛的和全球性的信息服务中心，其中包含了丰富的超链接信息，为数据挖掘提供了丰富的资源。
- **Web 数据挖掘包括 Web 使用模式挖掘、Web 结构挖掘和 Web 内容挖掘等。**





# 常用的数据挖掘方法

## 1. 分类与预测

分类和预测是两种重要的数据分析方法，在商业上的应用很多。分类和预测可以用于提取描述重要数据类型或预测未来的数据趋势。

分类是找出一个类别的概念描述，它代表了这类数据的整体信息，即该类的内涵描述。一般用规则或决策树模式表示。该模式能把数据库中的元组影射到给定类别中的某一个。

预测是利用历史数据找出变化规律，建立模型，并用此模型来预测未来数据的种类，特征不等。典型的方法是回归分析，即利用大量的历史数据，以时间为变量建立线性或非线性回归方程。

分类的方法主要有：**决策树 (C5 或 CART)、贝叶斯分类、基于遗传算法分类**

预测的方法主要是回归统计，包括：**线性回归、非线性回归、多元回归、泊松回归、对数回归等**。分类也可以用来预测。**神经网络方法预测**既可用于连续数值，也可以用于离散数值。

## 2. 关联分析

**关联分析** -- 就是挖掘数据对象之间的相互依赖关系。

**关联** -- 若两个或多个变量的取值之间存在某种规律性，就称为关联。

一个关联规则的形式为：

$$A1 \wedge A2 \wedge \cdots \wedge Ai \rightarrow B1 \wedge B2 \wedge \cdots \wedge Bj$$

其含义为：如果  $A1 \wedge A2 \wedge \cdots \wedge Ai$ ，则一定出现  $B1 \wedge B2 \wedge \cdots \wedge Bj$

**数据中的关联可分为：**

- **简单关联**

如：买面包的顾客中有 90% 的人购买了牛奶。面包 → 牛奶

- **时序关联**

如：粮食涨价，不久副食品涨价。

- **因果关联**

属条件与结论的依赖关系。

## ■ 聚类分析

■ 将数据点分组的过程，从而使得同一组内的数据点类似。

- 检查一大群最初没有差异的顾客，看看能否把它们分在自然形成的组内。

■ 聚类不同于分类的区别在于结果是分析出来的而不是事先预定的。

- 没有预先制定的设想，希望数据挖掘工具能够揭示某些有意义的结构。

聚类技术主要包括：模式识别方法、数学分类法、概念聚类、神经网络的自组织模型等。

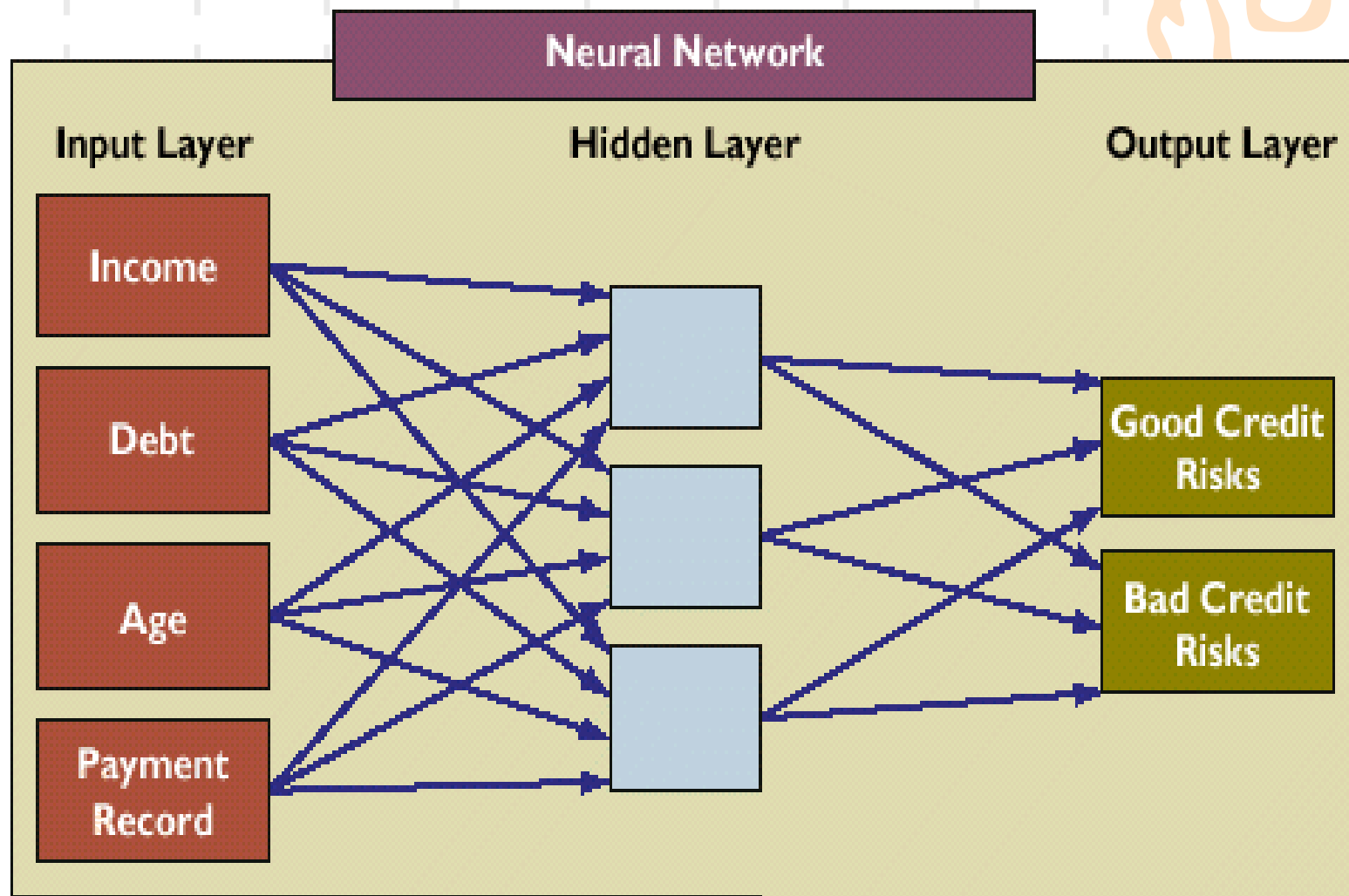
## 4. 人工神经网络

- 人工神经网络，是对人类大脑系统机能的简单抽象和模拟；
- 神经网络是一组连接的输入 / 输出单元，其中每个连接都与一个权相关联，在学习阶段，通过调整神经网络的权，使得能够预测输入样本的正确类标号来学习。
- 具有高度抗干扰能力和可以对未训练的数据分类的特点
- 激励函数的选择和权值的调整

将人工神经网络应用于数据挖掘的主要缺点是，通过人工神经网络学习到的知识难于理解；学习时间太长，不适于大型数据集。

# 神经网络

吉祥如意



吉祥如意

吉祥如意

吉祥如意

## 5 . 偏差检测

对数据库中的异常数据进行检测，称为偏差检测。

**偏差检测的基本方法**：寻找观察结果与参照之间的差别。

**观察**：通常是某一个域的值或多个域值的汇总。

**参照**：是给定模型的预测、外界提供的标准量或另一个观察。

**偏差检测的数据模式**有：极值点、断点、拐点、零点和边界等不同的偏差对象。

**偏差包括的规则知识**有：分类中的反常实例；模式的例外；观察结果对模型预测的偏差；量值随时间的变化等。

# 数据挖掘常用的 10 大算法

## 1. C4.5、C5.0 算法：

- C4.5 算法是机器学习算法中的一种分类决策树算法，其核心算法是 ID3 算法。C4.5 算法继承了 ID3 算法的优点，并在以下几方面对 ID3 算法进行了改进：
  - 1) 用信息增益率来选择属性，克服了用信息增益选择属性时偏向选择取值多的属性的不足；
  - 2) 在树构造过程中进行剪枝；
  - 3) 能够完成对连续属性的离散化处理；
  - 4) 能够对不完整数据进行处理。
- 优点：产生的分类规则易于理解，准确率较高。
- 缺点：在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而导致算法的低效。



## 2. K-Means 算法

- **k-means algorithm** 算法是一个聚类算法，把  $n$  的对象根据他们的属性分为  $k$  个分割， $k < n$ 。它与处理混合正态分布的最期望算法很相似，因为他们都试图找到数据中自然聚类的中心。它假设对象属性来自于空间向量，并且目标是使各个群组内部的均方误差总和最小。即每个簇用该簇中对象的平均值来表示。

## 3. Support vector machines

- 支持向量机，英文为 **Support Vector Machine**，简称 **SV** 机（论文中一般简称 **SVM**）。它是一种监督式学习的方法，它广泛的应用于统计分类以及回归分析中。支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔的超平面。在分开数据的超平面的两边建立有两个互相平行的超平面。分隔超平面使两个平行超平面的距离最大化。假定平行超平面间的距离或差距越大，分类器的总误差越小。

## 4. 经典的 Apriori 算法

---

算法思想：Apriori 算法思想基于如下定理：

若  $c[k] \in$  频繁集,  $m < k$   $c[m] \subset c[k]$

则  $c[m] \in$  频繁集

故可以用短的频繁集中元素构造长的频繁集  
元素

算法目的：提高频繁集发现效率

## ❖ 5. 最大期望 (EM) 算法

❖ 在统计计算中，最大期望（EM，Expectation–Maximization）算法是在概率（probabilistic）模型中寻找参数最大似然估计的算法，其中概率模型依赖于无法观测的隐藏变量（Latent Variable）。最大期望经常用在机器学习和计算机视觉的数据集聚（Data Clustering）领域。

## ❖ 6. PageRank

❖ PageRank 是 Google 算法的重要内容。2001 年 9 月被授予美国专利，专利人是 Google 创始人之一拉里·佩奇（Larry Page）。因此，PageRank 里的 page 不是指网页，而是指佩奇，即这个等级方法是以佩奇来命名的。

## ❖ 7. Naive Bayes

- ❖ 假定一个属性值对给定类的影响独立于其他属性的值
- ❖ 在众多的分类模型中，应用最为广泛的两种分类模型是决策树模型 (Decision Tree Model) 和朴素贝叶斯模型 (Naive Bayesian Model, NBC)。朴素贝叶斯模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。

## ❖ 8. CART: 分类与回归树

- ❖ CART, Classification and Regression Trees。

- ❖ 算法采用一种二分递归分割的技术，将当前的样本集分为两个子样本集，使得生成的决策树的每个非叶子节点都有两个分支。因此，CART 算法生成的决策树是结构简洁的二叉树。在分类树下面有两个关键的思想。第一个是关于递归地划分自变量空间的想法；第二个想法是用验证数据进行剪枝。

## ❖ 9.kNN: k-nearest neighbor classification

- ❖ K 最近邻 (k-Nearest Neighbor, KNN) 分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：如果一个样本在特征空间中的  $k$  个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别。

## ❖ 10.AdaBoost

- ❖ Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器（弱分类器），然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。

# 数据挖掘工具简介

目前，世界上比较有影响的典型数据挖掘系统包括：

- Enterprise Miner ( SAS 公司)
- Intelligent Miner ( IBM 公司)
- SetMiner ( SGI 公司)
- Clementine ( SPSS 公司)
- Warehouse Studio ( Sybase 公司)
- See5 ( RuleQuest Research 公司)
- CoverStory
- EXPLORA
- Knowledge Discovery Workbench
- DBMiner
- Quest 等

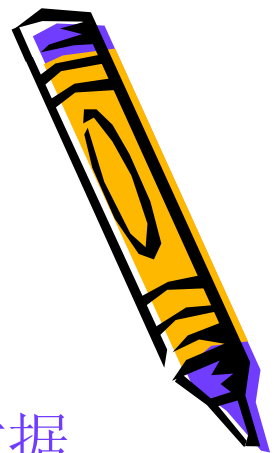


# 三大工具总体评分

功能	总分			
	权值	软件		
		IBM	SAS	SPSS
		Intelligent Miner	Enterprise Miner	Clementine
数据存取	10%	75	90	80
数据处理	20%	93	100	98
模型算法	30%	91	96	91
自动建模	10%	92	100	86
可视化	15%	88	95	91
其它	15%	78	92	56
总分	100%	88	96	86



# 数据挖掘工具介绍— Intelligent Miner



- 美国 **IBM** 公司开发的数据挖掘软件，分别面向数据库和文本信息进行数据挖掘的，包括 **Intelligent Miner for Data** 和 **Intelligent Miner for Text** 。
- **Intelligent Miner for Data** 可以挖掘包含在数据库、数据仓库和数据中心中的隐含信息，帮助用户利用传统数据库或普通文件中的结构化数据进行数据挖掘。已经成功应用于市场分析、诈骗行为监测及客户联系管理等；
- **Intelligent Miner for Text** 允许企业从文本信息进行数据挖掘，文本数据源可以是文本文件、 **Web** 页面、电子邮件、 **Lotus Notes** 数据库等等。



# 数据挖掘工具介绍— SAS Enterprise Miner

**SAS** 是一个庞大的系统，它多个功能模块组成，每个模块分别完成不同的功能。由于 **SAS** 最初是为专业统计人员设计的（这一点和 **SPSS** 已恰恰相反），因此使用上以编程为主。

## SEMMA 方法

**Sample**— 数据取样（质量、目标）

**Explore**— 数据特征探索、分析和预处理

**Modify**— 问题明确化、数据调整和技术选择

**Model**— 模型的研发、知识的发现

**Assess**— 模型和知识的综合解释和评价

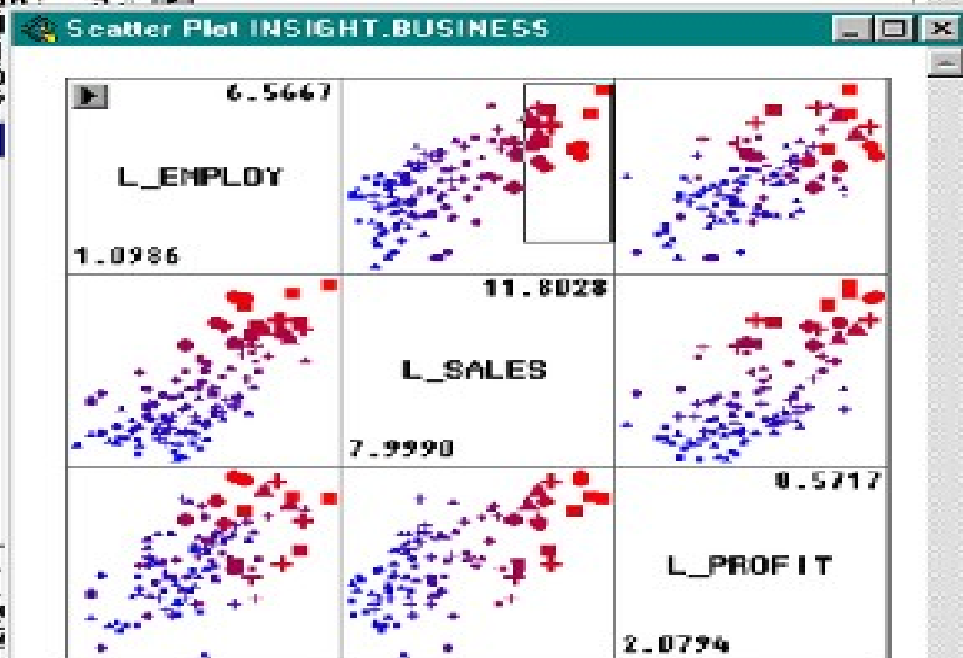
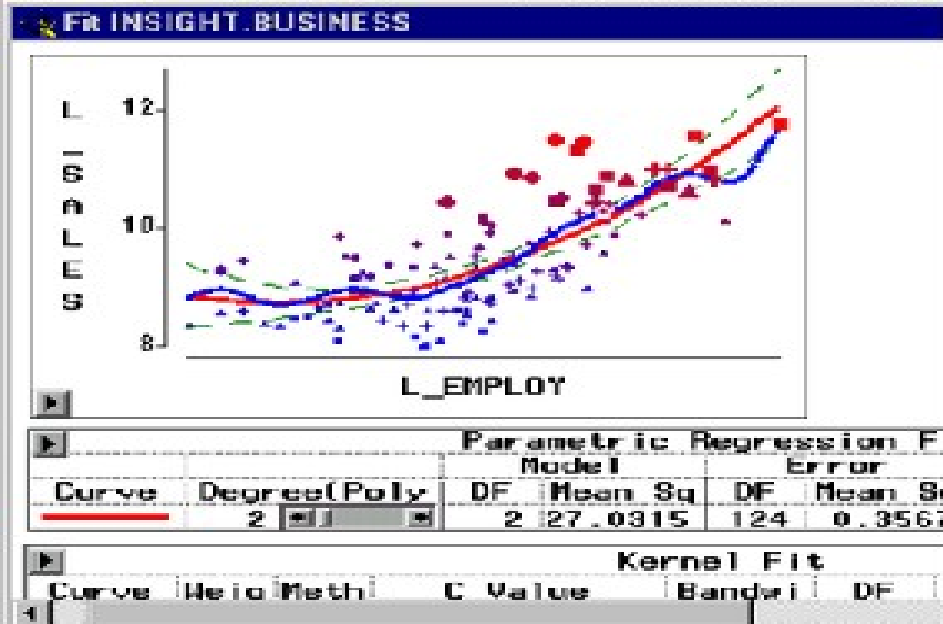
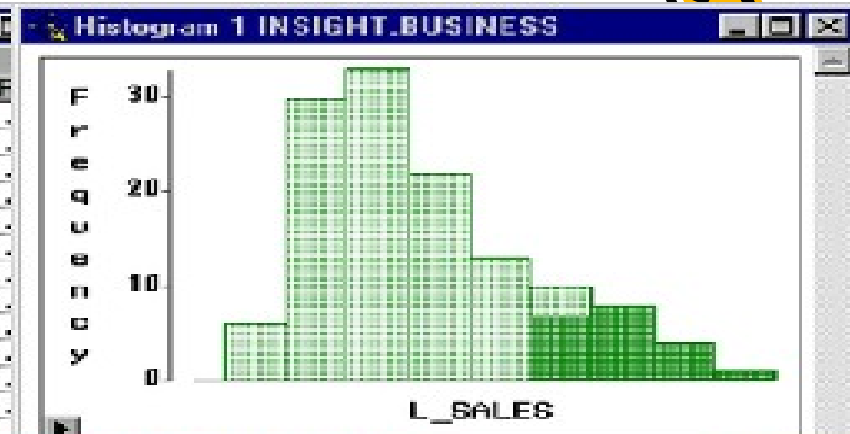


# SAS Enterprise Miner: scatter plots



INSIGHT.BUSINESS

		Int	Int	Int	Int	
127		SALES	PROFITS	L_EMPLOY	L_SALES	L_PROFIT
39		\$9,414	\$-236	2.6391	9.1500	
40		\$9,491	\$907	4.5326	9.1501	6.
41		\$3,037	\$58	3.2958	8.0186	4.
42		\$60,823	\$4,318	5.4027	11.0157	8.
43		\$8,135	\$506	4.7958	9.0039	6.
44		\$133,622	\$2,468	6.5667	11.8028	7.
45		\$11,164	\$629	4.4659	9.3204	6.
46		\$7,006	\$650	3.1355	8.8545	6.
47		\$7,103	\$396	3.6376	8.8683	5.
48		\$3,319	\$91	3.4340	8.1074	4.
49		\$6,900	\$142	3.7612	8.8393	4.
50		\$4,963	\$24	2.1972	8.5090	3.
51		\$95,790	\$220	4.5109	10.4056	5.
52		\$11,671	\$90	3.3673	9.364	
53		\$12,857	\$11	1.6094	9.461	
54		\$8,782	\$2,298	3.4012	9.080	
55		\$13,731	\$-38	2.5649	9.527	



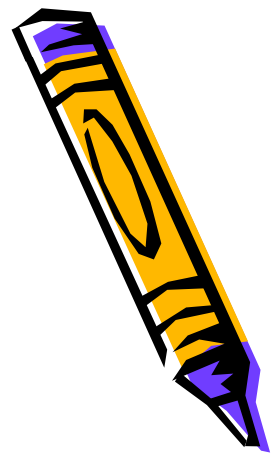
# 数据挖掘工具介绍— IBM Intelligent Miner



- 美国 **IBM** 公司开发的数据挖掘软件，分别面向数据库和文本信息进行数据挖掘的，包括 **Intelligent Miner for Data** 和 **Intelligent Miner for Text** 。
- **Intelligent Miner for Data** 可以挖掘包含在数据库、数据仓库和数据中心中的隐含信息，帮助用户利用传统数据库或普通文件中的结构化数据进行数据挖掘。已经成功应用于市场分析、诈骗行为监测及客户联系管理等；
- **Intelligent Miner for Text** 允许企业从文本信息进行数据挖掘，文本数据源可以是文本文件、**Web** 页面、电子邮件、**Lotus Notes** 数据库等等。



# 数据挖掘工具介绍— Spss 的 Clementine



- Clementine 是 ISL(Integral Solutions Limited) 公司开发的数据挖掘工具平台。1999 年 SPSS 公司收购了 ISL 公司，对 Clementine 产品进行重新整合和开发。
- 是一个开放式数据挖掘工具，曾两次获得英国政府 SMART 创新奖。
- 不但支持整个数据挖掘流程，从数据获取、转化、建模、评估到最终部署的全部过程，还支持数据挖掘的行业标准 --CRISP-DM。





## • 主要功能

分类：类神经网络、决策树 (C5 或 CART)、Logistic 回归；

聚类：K-Means 算法 (一维聚类)、Kohonen 算法 (利用类神经网络自我组织的演算法进行二维聚类)、2-Step 算法 (可自动找出最适合的聚类数)；

关联：Apriori 算法 (连续、类别变量都可用)、GRI 算法 (只能处理类别变量)、序列算法 (只能处理类别变量，且考虑时间先后)。



## ● Clementine 数据源

ODBC( 包括 Excel)

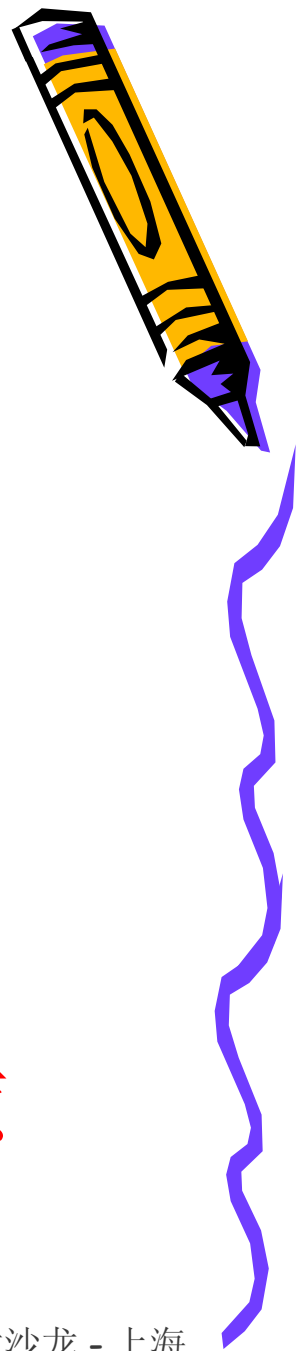
各种文本文件

Spss 数据源

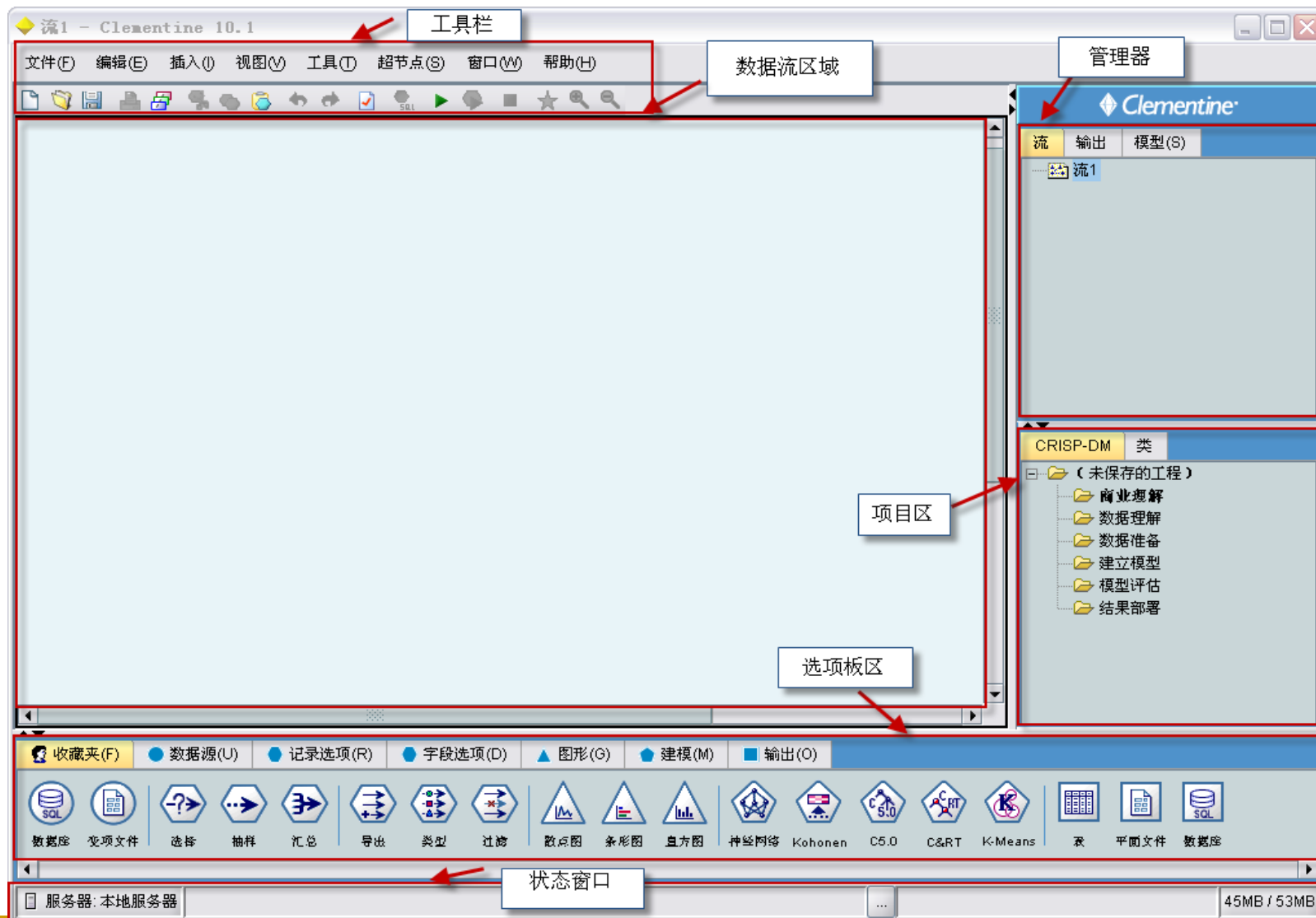
SAS 数据源

使用者输入

## ● Clementine 可同时存取多种数据来源



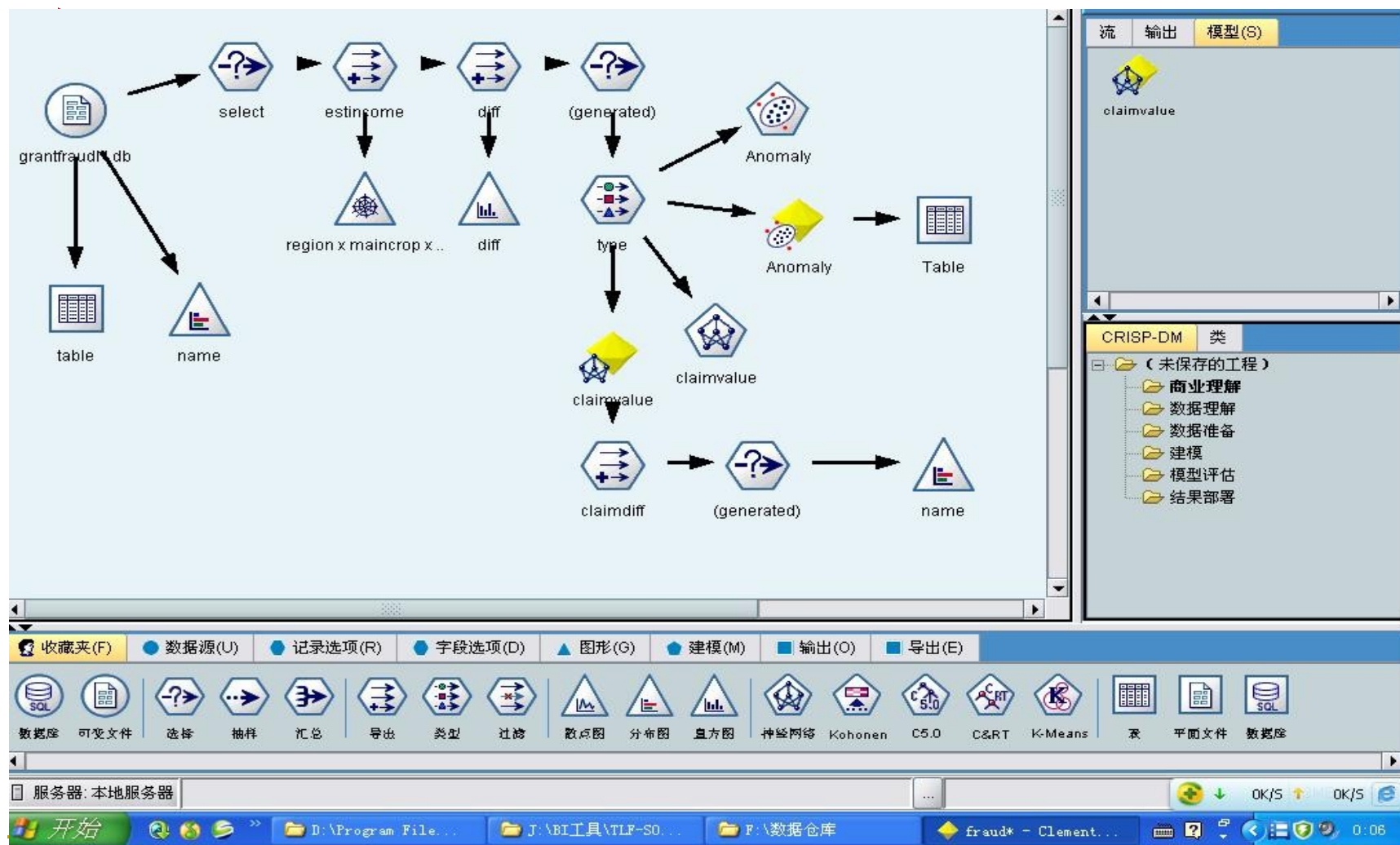
# Clementine 操作界面





# 筛选潜在诈欺案例的两种方法

## - “异常检测”和基于“神经网络”的建模方



# 数据挖掘技术应用广泛

- 数据挖掘技术从一开始就是面向应用的。由于现在各行业的业务操作都向着流程自动化的方向发展，企业内产生了大量的业务数据。
- 数据挖掘技术应用很广，应用较好的领域有：
  - 金融保险业： Credit Scoring ; Insurance Evaluation
  - 电信： Detecting telephone fraud
  - 零售（如超级市场）等商业领域： Marketing Analysis
  - 医学： Detecting inappropriate medical treatment
  - 体育： IBM Advanced Scout analyzed NBA game statistics
  - 在天文学、分子生物学等科学研究方面
  - 军事方面：使用 DM 进行军事信息系统中的目标特征提取、态势关联规则挖掘等。

# 市场营销的应用

- ❖ 基于购买模型分析顾客行为；
- ❖ 识别顾客流失模型以及通过预防行为使顾客未流失的情况；
- ❖ 广告、仓库位置等营销战略的确定；
- ❖ 顾客、产品、仓库的划分；
- ❖ 目录设计、仓库布局、广告活动；
- ❖ 通过适当聚集和为前端销售、服务人员发送信息，提供优先销售和顾客服务；
- ❖ 鉴定市场高于或低于平均增长；
- ❖ 识别同时被购买的产品，或购买某种产品类别的顾客特征；
- ❖ 市场容量分析。

# 财务的应用

- ❖ 客户信誉价值分析；
- ❖ 帐户应收款项划分；
- ❖ 金融投资，如股票、共有基金、债券等的业绩分析；
- ❖ 风险评估和欺诈检测

# 制造业的应用

- ❖ 优化资源，例如人力、机器、材料、能量等等；
- ❖ 优化制造过程设计；
- ❖ 产品设计；
- ❖ 发现生产问题的起因；
- ❖ 识别产品和服务的使用模型。

# 银行业务的应用

- ❖ 检测欺诈性信用卡使用的模型；
- ❖ 识别忠实顾客；
- ❖ 预测可能改变他们的信用卡从属关系的客户；
- ❖ 确定客户群体的信用卡消费。



# 医疗保健的应用

- ❖ 发现放射线图象的模型；
- ❖ 分析药物的副作用；
- ❖ 描述患者行为特征，预测外科手术观察；
- ❖ 标识对不同疾病的成功药物疗法。

## ❖ 竞技运动中的数据挖掘

大约 20 个 NBA 球队使用了 IBM 公司开发的数据挖掘应用软件 **Advanced Scout** 系统来优化他们的战术组合。

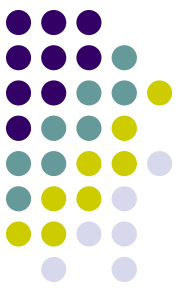
例如 **Scout** 就因为研究了魔术队队员不同的布阵安排，在与迈阿密热火队的比赛中找到了获胜的机会。

---- 系统分析显示魔术队先发阵容中的两个后卫安佛尼·哈德卫 (**Anfernee Hardaway**) 和伯兰·绍 (**Brian Shaw**) 在前两场中被评为 - 17 分，这意味着他俩在场上，本队输掉的分数比得到的分数多 17 分。然而，当哈德卫与替补后卫达利尔·阿姆斯创 (**Darrell Armstrong**) 组合时，魔术队得分为正 14 分。



# 刑事案件中的应用

- ❖ 三联生活周刊的报道
- ❖ 图森的一起谋杀案：一个男人被人切断了喉管，并被汽车碾过身体。当被发现时，他依然活着，并在被送往医院前告诉现场围观者——“这是‘矮子’干的”。
- ❖ 警方将“矮子”这个名字输入到 **Coplink** 数据库中，搜索它与被害人的联系。几分钟之内，**Coplink** 就给出了结果：被害人曾经与这个“矮子”共同在监狱中服刑。



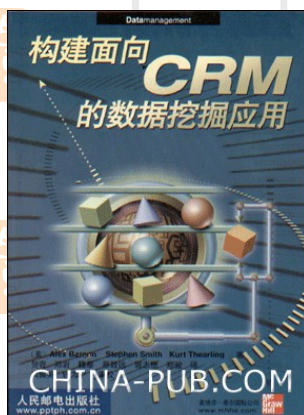
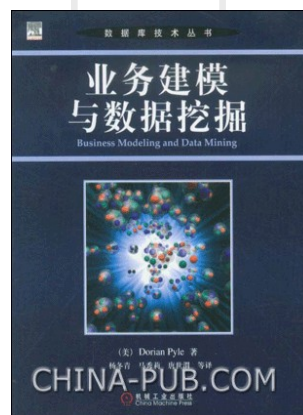
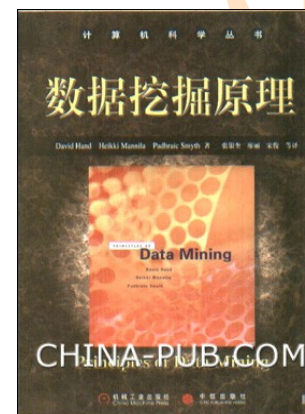
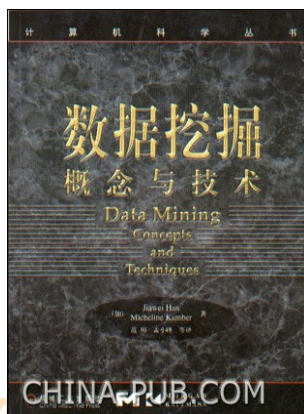
# 数据挖掘的前景

- 就目前来看，将来研究的几个焦点包括：
  - 研究在网络环境下的数据挖掘技术（**WebMining**），特别是在因特网上建立 **DMKD** 服务器，并且与数据库服务器配合，实现分布式数据采掘；
  - 生物信息或基因（**Bioinformatics/genomics**）的数据挖掘
  - 加强对各种非结构化数据的开采（**DataMiningforAudio & Video**），如对文本数据、图形数据、视频图像数据、声音数据乃至综合多媒体数据的开采；
  - 寻求数据挖掘过程中的可视化方法，使知识发现的过程能够被用户理解，也便于在知识发现的过程中进行人机交互；
  - 处理的数据将会涉及到更多的数据类型，这些数据类型或者比较复杂，或者是结构比较独特。
  - 发现语言的形式化描述，即研究专门用于知识发现的数据挖掘语言，也许会像 **SQL** 语言一样走向形式化和标准化；

# 几点体会总结

- ▶ 数据挖掘是年轻充满希望的研究领域
- ▶ 实施数据挖掘是一个战略性举措
- ▶ 数据挖掘是一个循环探索的过程
- ▶ 数据挖掘不是万能的解决方案

# 参考文献



商业智能研讨沙龙 - 上海站  
ITPUB ChinaUnix IX PUB 主办

# 网络资源



◆ [www.dwway.com](http://www.dwway.com)



◆ [www.dmresearch.net](http://www.dmresearch.net)



◆ [www.dmreview.com](http://www.dmreview.com)



◆ [www.kdnuggets.com](http://www.kdnuggets.com)



◆ [www.datawarehouse.com](http://www.datawarehouse.com)

商业智能研讨沙龙 - 上海站  
ITPUB ChinaUnix IXPUB 主办





# Thank You !